MS26 Communications PoC between NAKALA and HAL

Adeline Joffres, Nicolas Larrousse, Yannick Barborini, Bénédicte Kuntziger-Planche

MS26 Communications PoC between NAKALA and HAL

	PoC: Working Communication between Nakala, the Huma-Num's research data repository, and the academic publications available on HAL.
Lead Partner:	CNRS
Version:	**V1**
Status:	**Work In Progress**
Dissemination Level:	**All the project**
Document Link:	PDF version (linked to Participant Portal): https://repository.eosc-pillar.eu/index.php/s/W3EBLpFopHprGZ4

Document Abstract

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 857650.

This document specifies the development of the proof of concept (POC) which will allow the linking of publications deposited in HAL (including its HAL-SHS portal), the French open archive (https://hal.archives-ouvertes.fr/) developed by the CCSD (https://www.ccsd.cnrs.fr/), and data or datasets deposited in Nakala (https://nakala.fr/) developed by Huma-num (https://www.huma-num.fr/), the French research data repository for SSH data.

It describes the workflow to be set up between the two repositories, the creation of the bidirectional relationship between publications and data, prospection for the implementation of tools to visualize these links in the two repositories, and finally how to expose these relationships and publish them, so that they are available, findable and usable by research communities (and by all citizens).

DELIVERY SLIP

	Name	Partner/Activity	Date
From:	Adeline Joffres,	CNRS (Huma-Num)	14/12/2020
	Nicolas Larrousse	CNRS (Huma-Num)	14/12/2020
	Yannick Barborini	CCSD	14/12/2020
	Bénédicte Kuntziger	CCSD	14/12/2020
Reviewed by:			
Approved by:			

DOCUMENT LOG

Issue	Date	Comment	Author
V1.0	17/12/2020	This is the version shared on the Participant Portal.	

TERMINOLOGY

https://eosc-portal.eu/glossary

Terminology/Acronym	Definition
DOI	Digital Object Identifier
EOSC	European Open Science Cloud
FAIR	Findable Accessible Interoperable Reusable
PID	Persistent Identifier
POC	Proof Of Concept

Contents

Executive summary

This deliverable specifies the development of the proof of concept (POC) which will allow the linking of publications deposited in HAL (including its HAL-SHS portal), the French open archive (https://hal.archives-ouvertes.fr/) developed by the CCSD (https://www.ccsd.cnrs.fr/), and data or datasets deposited in Nakala (https://nakala.fr/) developed by Huma-num (https://www.huma-num.fr/), the French research data repository for SSH data.

It describes the workflow to be set up between the two repositories, the creation of the bidirectional relationship between publications and data, prospection for the implementation of tools to visualize these links in the two repositories, and finally how to expose these relationships and publish them, so that they are available, findable and usable by research communities (and by all citizens).

1 Introduction

The new publishing models combine articles and underlying data, deposited in publisher-specific or publisher-recommended repositories, which contributes to the transparency and reproducibility of research, allows for better evaluation of results and promotes data reuse.

But beyond the open access publishing model, open science principles require that publications be deposited in open archives and that data be managed according to FAIR principles, and deposited in data repositories. Today, there are different types of repositories, publication repositories and data repositories, which can be generalist, disciplinary or institutional, independently of each other, and without the deposited objects being linked to each other. The search for, and the exploitation of, these research objects are therefore decorrelated. Linking the different objects together, publications, data, as well as the software codes produced during a research project makes it possible to put the research in context, to increase its visibility and its reusability.

The aim of this POC is thus to increase interoperability between repositories. The objective is to create the relationships between the publications deposited in the HAL open archive, and the data deposited in the Nakala data repository, and to expose the relationships thus created so that they can be easily found and exploited. A second objective is to test tools for displaying and disseminating these links, so that they can be exploited in the European Web of data (e.g. Scholexplorer https://scholexplorer.openaire.eu/, or B2Note https://e-sdf.github.io/b2note-docs).

2 Description of the two repositories

2.1 HAL

The HAL open archive (https://hal.archives-ouvertes.fr) is the common platform, shared by the French academic community, for the open access dissemination of scientific production.

HAL provides access to the full-text of journal articles, or papers, theses, reports and so on. The document may or may not have been published, so both journal articles and preprints can be found in the archive. Documents from all academic fields can be submitted to HAL. The humanities and social sciences represent 20 percent of the documents (files) submitted to HAL.

HAL is a central repository for infrastructure (technical centralization), but is multifaceted (distributed portals), based on the metadata describing each deposit. One of these portals is HAL-SHS, which shows archives and disseminates scientific literature in all human and social science disciplines. HAL provides:

- an OAI infrastructure ensuring interoperability (OAI-PMH), API and Sword protocol
- dissemination of metadata through a Triple Store
- assignment of a PID (Persistent IDentifier), making data and metadata citable
- stability of identifiers (URL of deposits in particular)
- preservation of documents thanks to a partnership with CINES for archiving
- scientific quality of documents deposited as well as the details describing them: all deposited documents are therefore checked before being put online
- time stamping of deposits guaranteeing the intellectual property rights of the text deposited
- authority files available via the AURéHAL platform, which is interoperable with other listings.

HAL has been developing relationships and links with other international repositories, such as ArXiv, PubMedCentral and Software Heritage. When submitting a document or software code in HAL, the researcher can ask to push it towards one of these repositories, and HAL retrieves the repository identifier and displays it in the identifiers metadata of HAL.

In HAL, there are already different ways to create a relationship between a publication and another resource:

- A specific interface for linking documents together.
    - This functionality is only available for documents within HAL (with possible typing of the relationship). It requires being connected to HAL and having rights on the repository (property of the repository).
        - Cf. appendix for the types of relations proposed in HAL.

- Specific metadata at the time of the deposit: several fields make it possible to link the deposit to another resource:
  - Other identifier: corresponds to the identifier of the document in another repository (arxiv, pubmed, ...)
  - See also: URI to a document associated with the publication
  - Associated data: DOI to data (or a dataset) deposited in another repository
  - Linked publications: identifier of the associated publication (field accessible for software and data deposit (MediHal).
- Retrieval of bibliographical references associated with the publication using the GROBID tool (https://grobid.readthedocs.io/en/latest/Introduction/).

These various links are possible today, but are not exploited or are underexploited. Relationships can be added in several places: in the repository (link to a repository outside HAL) and as a feature outside the repository (specific interface to link repositories in HAL together).

Characteristics of the Relationships: the entry is manual in both cases (DOI or URL linked in the publication metadata / establishment of a relationship to link two objects in HAL).

2.2 NAKALA

NAKALA is an interoperable and secure service for depositing all types of data (text files, audio, video, images etc.) in order to share them. Based on OpenSource technologies such as Symfony, Elasticsearch, React, etc. this repository mainly provides different types of services:

- assignment of a PID (Persistent IDentifier) making data and metadata citable;
- permanent data access through Web frontend, an integrated search engine and APIs;
- dissemination of metadata through a Triple Store and OAI-PMH;
- organization of data into collections;
- presentation of collections in specific websites through Nakala_Press.

The default format used by NAKALA is the DublinCore format qualified for metadata but more broadly any type of field of a specific format can be adopted.

Each data is visible on a dedicated page (landing page) which presents the visualization of the data (different viewers for common format files, IIIF, etc.) as well as the associated metadata. This page can be accessed via the DOI assigned to the data.

3 User stories

In order to best reflect the needs of researchers (user side) as well as the possibilities of implementing the link between data and publications (infrastructure side), we have developed the following use cases:

- An author wishes to link a publication already deposited in HAL with the data present in NAKALA.
- A person wishes to access/view the data associated with a publication.
- An author wishes to link data already stored in NAKALA with publications already published in HAL.
- A person wishes to access/view the publications associated to a data.
- As HAL manager, I would like to retrieve the relation(s) created in NAKALA (and vice versa)
- As a repository manager, I wish to make statistics on the citation of the data deposited in my repository.

4 Describing relationships within both repositories

4.1 Relationship between objects

Table 1 – Definition of the relationship between objects

Local PID	PID
Local repository	HAL or NAKALA for POC
Type of relationship	Type of relationship between objects. See the Appendix for what is already in place in some repositories
Remote identifier	Identifier
Remote repository	HAL or NAKALA for POC
Creation date	Date of creation of the relationship

The choice of the typology of the relationship will have to be made upstream of the developments, giving priority to interoperability. We will choose a vocabulary that already exists and evolves. The chosen vocabulary will be implemented in the repositories, used in the export formats, and will have to be made compatible with the vocabularies used in the data exposure formats (DC Terms, ScholeXplorer schema for example).

The licence granted to the relationship will be a CCO licence https://creativecommons.org/publicdomain/zero/1.0/deed.en, which allows full reuse.

4.2 Relationship building workflow

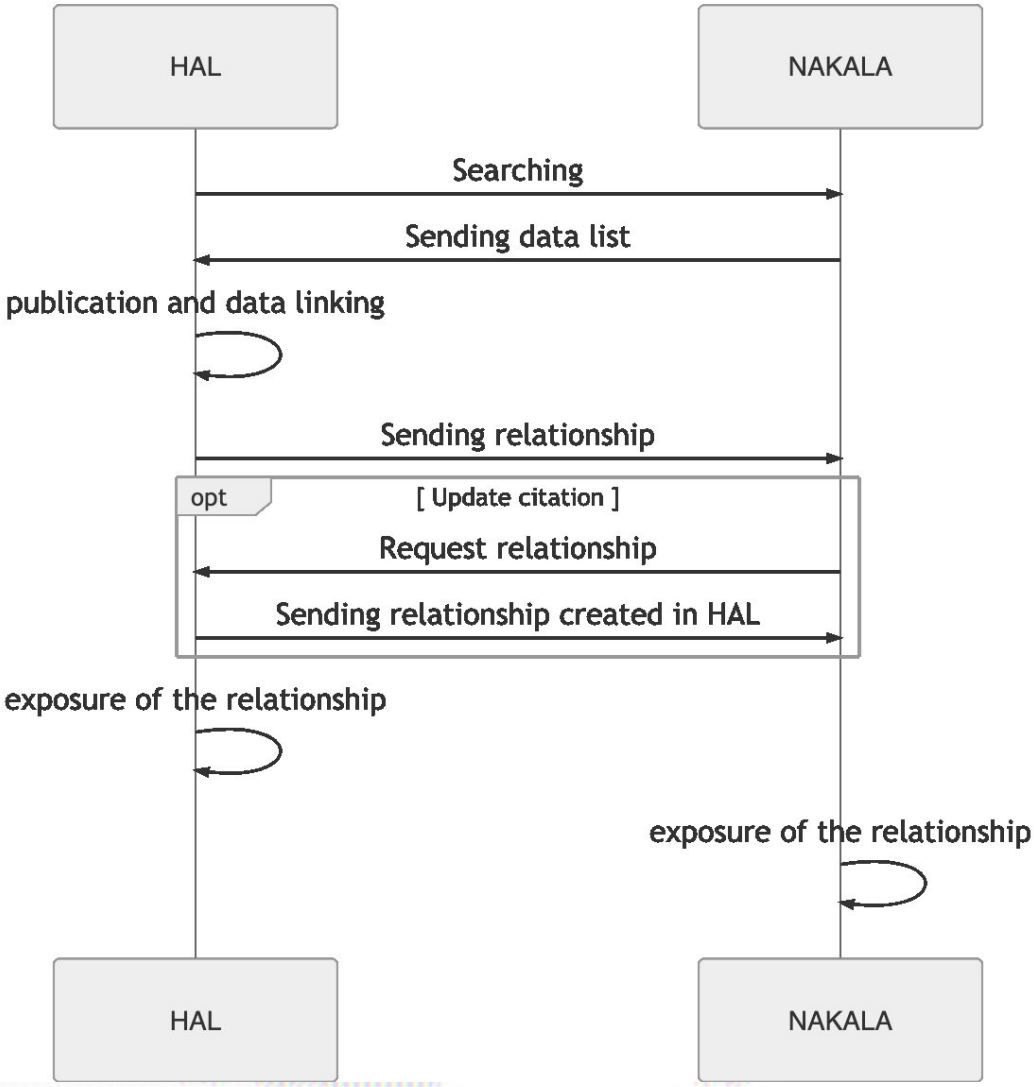

Fig.1 – Schema of the POC's workflow between HAL and NAKALA

The user can create a relationship between data deposited in HAL and data already stored in Nakala:

- When creating a new repository in HAL
- On a deposit already made in HAL

To be able to create a relationship, the user must have an account in HAL and be the owner of the data deposited in HAL for which he is creating a relationship (or have rights to edit the data). It is not necessary for the user to have rights to the data in NAKALA.

Searching for data in NAKALA:

- The user can indicate the data identifier (Handle or DOI).
- The user can search the data in Nakala thanks to a search API made available by NAKALA. The search will be carried out on the different metadata describing the data in NAKALA.

Table 2 – Search API

Description	Allows you to search in a repository. The search is carried out on the different metadata of the objects
Parameter	Text searched for (identifier or metadata title, author, ...)
Result	List of objects containing: • the identifier of the resource • the citation of the resource

This mechanism will be set up in both repositories in order to create relationships in HAL and NAKALA.

NB: In a first step, the applicant and users with the administrator role on the data are able to create the relationship in NAKALA.

5 Exposure of relationships

5.1 About data citation practices and links to publications in SSH and beyond

Although data citation has been a long-standing concern (see "Out of Cite, Out of Mind" - http://doi.org/10.2481/dsj.OSOM13-043), data citation practices in SHS are quite disparate and more recent even though the Social Sciences have developed considerable experience in this area.

This is due to the fact that research data were previously considered secondary. However, these practices are evolving rapidly, encouraged in particular by the need to establish data management plans which are required by agencies to obtain funding. In addition, data paper type objects are appearing in SSH (e.g. Cybergeo: European Journal of Geography has been publishing data papers since 2017: https://journals.openedition.org/cybergeo/28545), evidencing the interest in data today.

However, the types of data used by SSH are by nature very varied, which does not always make them easy to cite: the use of so-called dynamic data from social networks is a good example of this.

In brief, citing data especially in the context of SSH is important for different reasons:

- Providing a way to reproduce research which will in turn enhance the quality and effectiveness of research
- Reusing data for different research purposes in other contexts
- Giving credit to the creator and the funder of the data
- Proving the usefulness of infrastructures
- Enhancing links between data and publications

An inventory of current citation practices has been carried out within the framework of SSHOC (https://sshopencloud.eu/), a European SSH cluster project bringing together the main SSH actors: Inventory of SSH citation practices, and choice for SSHOC citation formats and implementation planning (https://doi.org/10.5281/zenodo.3595964).

The citation of data is generally based on various recommendations made in the framework of working groups, for example:

- APA & MLA-type citations that were originally intended for publications but have been adapted for datasets (See "MLA handbook for writers of research papers, New York: Modern Language Association of America, 2009, 7th ed.");
- The citation extension for electronic documents "ISO 690";
- RDAs Data Citation WG (https://www.rd-alliance.org/groups/data-citation-wg.html);
- Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk``

- The DataCite Metadata Working Group (See "DataCite Metadata Schema" https://schema.datacite.org/ and http://schema.datacite.org/meta/kernel-3/doc/DataCite- MetadataKernel_v3.1.pdf)

More recently, the creation of the notion of "Fair Digital Objects" (https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digi tal%20Objects/FDO-declaration) a synthesis of various previous studies (e.g. "A Framework for Distributed Digital Object Services" - http://www.cnri.reston.va.us/home/cstr/arch/k-w.html) will facilitate the citation of data even if the main purpose is to provide a framework for their interoperability.

The next step is to make these citations "actionable", i.e. machine-readable and machine-processable. The W3C has published recommendations for making data catalogues on the World Wide Web interoperable (https://www.w3.org/TR/vocab-dcat/).

Other services are being developed, for example by Crossref and Datacite to link datasets and other objects identified by DOIs. In this framework these institutions have developed a very comprehensive DOI citation formatting service (https://citation.crosscite.org/).

Large companies are also increasingly present on these subjects. One cannot fail to mention the "Google Data Search" service which has recently gone into official production (https://datasetsearch.research.google.com/), and traditional publishing houses are also positioning themselves such as Elsevier with its service for Mendeley data (Mendeley Data - https://data.mendeley.com).

It is therefore crucial today to have tools for linking data and publications for the trusted repositories offered to communities.

5.2 Repositories web interface

In HAL, the relationships created will be displayed on the landing page of a publication.

The citation of the data will be displayed, making it possible to bounce on the landing page of the data in NAKALA in order to have a more precise description of the data.

5.3 APIs

Table 3 - Search API description

Search API	
Description	Allows you to retrieve the citation of a data or publication.
Parameter	Identifier (DOI, Handle, HALID)
Result	Citation of the resource

Implementation of an API enabling the retrieval of relations made in the other platform from an identifier. In this way a relation created in HAL can be retrieved in NAKALA and vice versa.

Table 4 - Relationship recovery API

Description	Allows you to retrieve the relationships created in a repository based on an identifier
Parameter	Identifier (DOI, Handle, HALID)
Result	List of relations

5.4 Exports in different formats

- Present in the different exports available in which it can be integrated (TEI, RDF, bibteX)
- Exhibited in HAL's Triple Store (RDF)
- Indexed in the HAL search engine (SOLR)
- Exhibited in HAL's OAI-PMH repository: DC terms

6 External tools to disseminate relationships

For the moment, we have considered the use of two different tools to disseminate the relationships created between HAL publications and NAKALA data.

6.1 B2Note

B2Note (https://www.eudat.eu/catalogue/B2NOTE) is a tool for making annotations on resources. The tool also allows the user to create relationships between resources. The relations created between HAL and NAKALA can be added in B2Note in order to be more widely disseminated.

- Documentation: https://e-sdf.github.io/b2note-docs
- Service : https://b2note.bsc.es/
- EOSC Technical specifications and interoperability guidelines[1] : annotation service EUDAT-B2NOTE ; https://wiki.eosc-hub.eu/display/EOSCDOC/Metadata+Management+and+Data+Discovery?preview=/68223176/68223179/EOSC_Technical_specification_AnnotationService_Proposal_v1.pdf

6.2 SCHOLIX, A Framework for Scholarly Link eXchange

The objective of the Scholix initiative is to establish a high-level interoperability framework for the exchange of information on the links between scientific literature and data.

It might be interesting to trace these relationships also in the Scholexplorer platform of OpenAire: OpenAire, as one of the Scholix hubs, assists in the global aggregation of data-literature link information.

- http://www.scholix.org/
- https://scholexplorer.openaire.eu/#/

[1] To support the seamless operation of services in the future European Open Science Cloud, the EOSC-hub has proposed a number of technical specifications and interoperability guidelines covering both common and federation services

7 Conclusion - Prospects

The first step is to build the relationship between the publications deposited in HAL and the data deposited in Nakala, using the APIs available in each of the repositories. The relations thus created will be displayed, exported and harvestable.

A second step will consist, if possible, in allowing the simultaneous deposit of publications and data in the same repository, HAL, and then transferring the data into Nakala, creating the relationship between the two repositories automatically, on the model of what has already been developed in HAL for software codes, in partnership with the Software Heritage repository.

It is also conceivable, by means of a survey of users of the two data repositories, to deepen the bi-directional character of the "simple" link initially created, as well as its visualization.

8 General Bibliography

Munafò, M., Nosek, B., Bishop, D. *et al.* A manifesto for reproducible science. *Nat Hum Behav* **1,** 0021 (2017). https://doi.org/10.1038/s41562-016-0021

Auer, Sören. (2018, January 22). Towards an Open Research Knowledge Graph (Version 1). Zenodo. http://doi.org/10.5281/zenodo.1157185

Yannick Barborini, Roberto Di Cosmo, Antoine R. Dumont, Morane Gruenpeter, Bruno Marmol, et al.. The creation of a new type of scientific deposit: Software. *RDA Eleventh Plenary Meeting, Berlin, Germany*, Mar 2018, Berlin, Germany. 2018. ⟨hal-01738741⟩

Roberto Di Cosmo, Morane Gruenpeter, Bruno Marmol, Alain Monteil, Laurent Romary, et al.. Curated Archiving of Research Software Artifacts: lessons learned from the French open archive (HAL). 2019. ⟨hal-02475835⟩

Burton, Adrian, & Koers, Hylke. (2016, March 31). ICSU-WDS & RDA Publishing Data Services WG Interoperability Framework Recommendations (Version 1.0). http://doi.org/10.15497/RDA00002

Burton, Adrian, Fenner, Martin, Haak, Wouter, & Manghi, Paolo. (2017, November 21). Scholix Metadata Schema for Exchange of Scholarly Communication Links (Version v3). Zenodo. http://doi.org/10.5281/zenodo.1120265

https://graph.openaire.eu/
Adrian Burton,Amir Aryani, Hylke Koers, Paolo Manghi et al. The Scholix Framework for Interoperability in Data-Literature Information Exchange *D-Lib Magazine, 23,1/2 (2017).* http://www.dlib.org/dlib/january17/burton/01burton.html

M.Y. Jaradeh,A. Oelen,K.E.Farfar,M. Prinz et al. Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge? K-CAP '19: Proceedings of the 10th International Conference on Knowledge Capture September 2019 Pages 243–246 https://doi.org/10.1145/3360901.3364435

Task Group on Data Citation Standards and Practices, C.-I., 2013. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal, 12, pp.CIDCR1–CIDCR7. http://doi.org/10.2481/dsj.OSOM13-043

Nicolas Larrousse, Daan Broeder, Jan Brase, Cesare Concordia, & Vasso Kalaitzi. (2019). SSHOC D3.2 Inventory of SSH citation practices, and choice for SSHOC citation formats and implementation planning (Version v1.0). Zenodo. https://doi.org/10.5281/zenodo.3595964

9 Appendices

9.1 Relationship typing

Depending on the repository they belong to, vocabularies are already available to define a relationship between documents.

Repository	Relations
HAL	• illustrate • is illustrated • requires • is required by • has part • is part of • references • is referenced by • has format • is format of • conforms to • has version • is version of
Scholix	https://zenodo.org/record/1120265#.Xy1HlhP7TUI • IsSupplementTo (indicates that A is a supplement to B when both are published together) • IsSupplementedBy (Indicates that B is a supplement to A when both are published together) • References (Indicates B is used as a source of information for A) • IsReferencedBy (Indicates A is used as a source of information by B) • IsRelatedTo (Indicates a generic relation between A and B)
Zenodo	https://developers.zenodo.org/#representation • cites this upload • is cited by this upload • is supplemented by this upload • is a supplement to this upload • is referenced by this upload • references this upload • is previous version of this upload • is new version of this upload • continues this upload • is continued by this upload • has this upload as part • is part of this upload • reviews this upload • is reviewed by this upload

	• documents this upload • is documented by this upload • is compiled/created by this upload • compiled/created this upload • is the source this upload is derived from • has this upload as its source • is identical to this upload • is an alternate identifier of this upload
Datacite	https://support.datacite.org/docs/relationtype_for_citation • IsCitedBy • Cites • IsSupplementTo • IsSupplementedBy • IsContinuedBy • Continues • HasMetadata • IsMetadataFor • IsNewVersionOf • IsPreviousVersionOf • IsPartOf • HasPart • IsReferencedBy • References • IsDocumentedBy • Documents • IsCompiledBy • Compiles • IsVariantFormOf • IsOriginalFormOf • IsIdenticalTo • IsReviewedBy • Reviews • IsDerivedFrom • IsSourceOf
OpenAire	https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/vocab_relationtype.html#vocab-relationtype-relationtype • IsCitedBy (indicates that B includes A in a citation) • Cites (indicates that A includes B in a citation) • IsSupplementTo (indicates that A is a supplement to B) • IsSupplementedBy (indicates that B is a supplement to A) • IsContinuedBy (indicates A is continued by the work B) • Continues (indicates A is a continuation of the work B) • IsDescribedBy (indicates A is described by B) • Describes (indicates A describes B) • HasMetadata (indicates resource A has additional metadata B) • IsMetadataFor (indicates additional metadata A for a resource B)

	• HasVersion (indicates A has a version B) • IsVersionOf (indicates A is a version of B) • IsNewVersionOf (indicates A is a new edition of B, where the new edition has been modified or updated) • IsPreviousVersionOf (indicates A is a previous edition of B) • IsPartOf (indicates A is a portion of B; may be used for elements of a series) • HasPart (indicates A includes the part B) • IsReferencedBy (indicates A is used as a source of - information by B) • References (indicates B is used as a source of information for A) • IsDocumentedBy (indicates B is documentation about/explaining A) • Documents (indicates A is documentation about/explaining B) • IsCompiledBy (indicates B is used to compile or create A) • Compiles (indicates B is the result of a compile or creation event using A) • IsVariantFormOf (indicates A is a variant or different form of B, e.g. calculated or calibrated form or different packaging) • IsOriginalFormOf (indicates A is the original form of B) • IsIdenticalTo (indicates that A is identical to B, for use when there is a need to register two separate instances of the same resource) • IsReviewedBy (indicates that A is reviewed by B) • Reviews (indicates that A is a review of B) • IsDerivedFrom (indicates B is a source upon which A is based) • IsSourceOf (indicates A is a source upon which B is based) • IsRequiredBy (indicates A is required by B) • Requires (indicates A requires B)